学校编码：10384　　分类号_____密级_____
学　　号：23220111153272　　UDC_____

厦门大学

硕士学位论文

金融文本理解的特征选择研究

Feature Selection Research on Financial Text Understanding

周喆

指导教师姓名：罗林开教授
专业名称：控制工程
论文提交日期：2014年5月
论文答辩时间：2014年5月
学位授予日期：2014年　月

答辩委员会主席：_____
评　阅　人：_____

2014年5月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：周赟

2014年 5月 17日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：周喆

2014年5月17日

摘要

互联网上海量的金融文本数据（财经新闻，研究报告，股吧论坛等）蕴含着丰富的信息，而这些信息与很多财经事件或证券走势存在着密切的关系。如何让机器自动理解这些大量的金融文本数据，给出有价值的信息，就成为一个很有价值的工作。

选择合适的特征词集合（特征集）是金融文本理解的一个基础且不可逾越的环节。合适的特征集不仅要有好的分类能力，而且还要有好的稳定性，即对训练数据的低敏感性。

本文首先根据金融文本理解的具体任务，初选一个较大的候选特征集。接着，应用 SVM-RFE 和随机森林特征选择算法从该候选特征集中选择特征，并分析其在稳定性方面的不足；然后，给出了一种基于训练样本摄动的组合特征选择算法，并比较了它与未组合的特征选择方法在特征选择的稳定性和分类性能上的优劣；实证结果显示基于组合的方法在取得了与未组合方法相近的分类准确率下，提高了特征子集的稳定性，且降低了分类准确率的标准差，说明了组合方法的有效性和优越性。此外，本文还对确定最优特征子集大小的方法进行研究，给出了一种基于训练准确率的特征子集大小确定算法。

关键词：金融文本理解；特征选择；组合

Abstract

A large number of financial text data on the Internet, i.e., financial news, research report and stock BBS, contain rich information which having close relationship with many financial events or stock price trends. How to make machine automatically understand these massive financial text data and providing people with valuable information has become a very valuable work.

Choose the appropriate set of feature words is a fundamental and insurmountable link of financial text data understanding. The appropriate set of feature words should not only have good classification ability, but also has a good stability(i.e., low sensitivity to variational train data.).

In this paper, roughing a large set of candidate features is selected roughly on the basis of the specific financial text comprehension tasks firstly. Then, apply SVM-RFE and the feature selection algorithm based on random forest to select features from the candidate set and analyze their shortages in stability; Then, a combined feature selection algorithm based on training data perturbation is proposed. We compares it with feature selection methods uncombined in the stability of feature selection and classification performance; The empirical results show that in the condition of that the combined method gets close classification accuracy with uncombined methods, the combined method improves the stability of feature selection and reduces the standard deviation of classification accuracy, showing the effectiveness and superiority of the combined method. In addition, this paper also analyses the methods determining the size of the optimal feature subset and proposes a feature subset size determination algorithm based on training accuracy.

Key words: financial text mining understanding; feature selection; combination

目录

CONTENTS

第一章 绪论

1.1 课题研究背景及意义

随着我国金融信息化的高速发展，金融信息量也得到了前所未有的增长。正如著名的经济学家，1978 年诺贝尔经济学奖获得者 Herbert A. Simon 所说，我们现在面临的“信息危机”是信息量过剩的问题，“信息社会中，没有组织和控制是我们不希望看到的，它反而会成为信息工作者的敌人”[1]，如何从浩瀚如海的金融网页上找到人们所需要的信息，如何帮助人们从激增的数据背后准确收集并选择感兴趣的信息，如何帮助决策者在日益更新和增加的爆炸信息中自动发现新的知识及它们之间的关系，已经成为信息技术和金融领域里重要的研究课题之一。

在大量金融数据产生和积累的过程中，人们对这些数据进行了一定的研究和分析，期望得到有价值的信息。以往的研究中从不同的角度来看，分为基本分析法[2]和技术分析法[2]。基本分析法是从影响市场变化的原因上来分析，着重于市场本身的内在价值上。影响市场变化的因素有很多，包括国际国内的金融形势和经济环境、相关经济政策和经济指标、其他经济部门各行各业的状况、上市公司行业地位、市场前景、财务状况等。而技术分析方法着重于市场行为，采用一定的技术手段从累积的大量历史数据（企业报表、证券市场成交额和成交量等）中找出蕴含的规律，来判断市场未来的变化趋势，给投资者做适当的指导信号。技术分析法对市场的反应直接、简单准确，可操作性强及其适用范围广的优点，使其对金融市场的研究趋于成熟。

如今，很多媒介信息例如财经网站、论坛、银行及证券公司网站等都蕴含着大量的信息：证券机构的研究报告、上市公司的财务报表、股票投资者的经验分享、股民的个人分析与预测、最新的时事热点等。这些信息对于投资者来说，能为投资者的决策作参考，还能及时反应投资者的感情倾向；对上市公司及企业的领导者来说，能促进他们的优胜劣汰；对证券市场监管者证监会来说，给了监管当局一定的参考信息，有利于监管当局对市场的管理并制定相应的政策规章。因

此，合理的利用这些信息，可为我们分析和理解金融市场提供一个全新的角度。

金融市场的变化日新月异，需要决策者对日益更新的媒介信息有绝对的洞察力和判断，做出有效恰当的反应，但是仅仅依靠人力不能充分利用丰富的信息资源，无法快速、准确、全面的得到分析结果，因此可以借助计算机技术为我们服务。这样我们就能节省时间、为研究提供便利。

媒介信息大多是非结构性的文本字样，文本挖掘作为一种特殊的数据挖掘[3]，可以从很多的、非结构化的文字样本中提取事先未知的、可以理解的、最终可用的知识[4]。在金融文本理解中，特征词的选择是个重要的环节，不同的特征词表征文本的结果也就不一样，例如对财经新闻来说，在标题为“股市大盘反弹将维持到周二”的新闻中，若文本内容中出现“上涨”、“反弹”、“阳线”等特征词，我们可以认为这篇新闻认定大盘趋势上行，这些特征词能更加帮助我们对文本进行深入的理解。在金融文本特征词方面，参考互联网上对股票趋势的专业术语，结合我国股票市场的实际情况，通常是根据经验初选一些特征词汇。

然而，初选的特征词汇通常具有较多的特征词，且很可能存在与金融文本理解任务相关度不高的词汇，因此需要对初选出来的特征选择通过一定的方法将其特征选择出来。因此需要采用特征选择方法对初选出的特征词汇进行特征选择。支持向量机递归特征消除(Support Vector Machine-Recursive Feature Elimination, SVM-RFE)算法[5-7]和随机森林特征选择算法[8,9]是广泛使用的算法，它们选出来的特征子集都拥有良好的分类功能。但是，这些方法训练出来的特征集都依赖于进行训练所需要的原数据，也就是说一旦原数据发生变化，它们选择出来的特征集可能会发生改变。但是，这些特征选择方法选出的特征子集是依赖于训练数据，也就是说一旦训练数据发生变化，它们选出的特征子集可能会发生改变。一般说来，特征子集应该是依赖于文本理解任务的，应该对特征选择方法和训练数据不敏感。因此，在金融文本理解的特征选择中，我们不仅要考虑选出的特征子集具有好的分类性能，而且还必须考虑它的稳定性问题，即特征子集对特征选择方法和训练数据的低敏感性。

为了获取拥有更好的分类性能并且稳定的特征集，结合组合方法可以提高分类的稳定性能，本文将 SVM-RFE 算法和随机森林特征选择算法应用于金融文本

挖掘的特征选择中，并在此基础上给出了一种基于训练样本摄动的组合特征选择算法，并比较了它与未组合的特征选择方法在特征选择的稳定性和分类性能上的优劣，尝试在金融文本的理解中，选出具有好的分类性能、又有好的稳定性能的特征词汇。

1.2 课题研究现状

1.2.1 文本挖掘的现状研究

目前，对短文本的挖掘逐渐发展成为国际上文本数据挖掘领域的新的热点。短文本的载体主要有博客、微博、论坛留言、在线评论、聊天记录等。短文本具有文本内容短小、存在大量的噪声、所含知识信号较弱、关键词特征稀疏、关键词维度高等特点。

国外已有许多学者对此做出了研究，上个世纪 50 年代末期，H.P.Luhn[10]率先提出了基于词频统计的方法，并在文本的自动分类中有了成功应用；Maron 和 Kuhns[11]等人在 1960 年第一次发表了和文本分类相关的文献；随后 M.E.Lesk、G.Salton、K.S.Jones [12-14]等学者在这一领域继续进行了具有一定意义的研究。到了 21 世纪，Pang [15]等人采用了多种方法对 Usenet 上的电影评论进行文本的情感分类，并和人工的分类结果进行比较；Turney[16]等人使用了点互信息法对产品评论进行了研究。目前，文本分类已从理论性研究向实用性应用转变，广泛应用在邮件分类、信息检索与过滤、数字图书馆、电子会议中。

文本自动分类大概经历了以下几个发展历程：

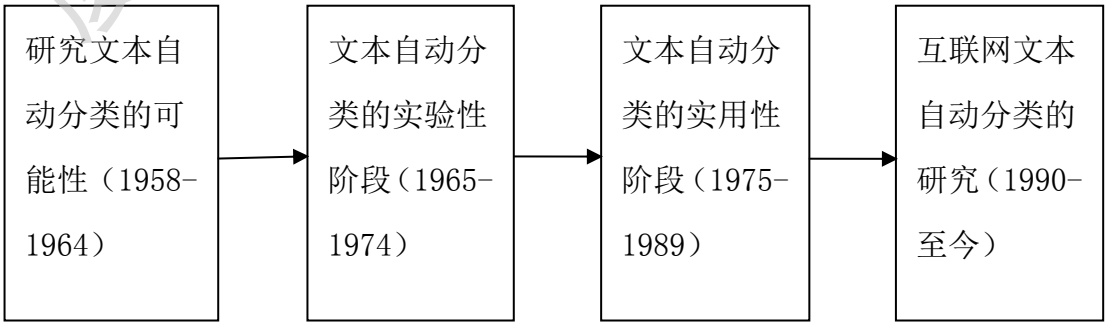

图 1.1 文本自动分类的发展历程

国内学者对 PC 在文本分类工作当中进行研究的是清华大学的侯汉清教授[17]，他还对通过 PC 进行分类表的管理、分类编制、分类检索等方面做了概括；闫瑞[18]等学者提出了通过动态调整的策略进行组合分类器的训练，在中文聊天记录分类中获得了更好的效果；林小俊[19]等人建议的一种以概念网络为基础的短文本分类这种方法，进一步让档案领域的短文本的分类成为现实。与此同时，国内自动分类系统和计算机辅助分类系统也得到了发展，相比于国外卡内基集团为法国路透社的上千稿件进行自动分类的 CONSTRUE 系统[20]，因为中文和英文这两种语言间存在很大的差异，研究人员无法直接使用国外的研究成果，所有国内多所高校及研究单位对中文文本挖掘也有了自己的研究成果，例如李晓黎、石忠植教授[21]应用概念推理网进行的文本分类研究，返回率可达到 87.2%，准确率达到 89.4%；吴立德等人[22]研究的独立语种的文本分类，分析了单分类和多分类等方面，并且通过词汇的种类不同和信息量的大小作为判断标准，最好的返回率达到 83.87%，以及清华大学研究的自动分类系统都具有良好的效果。目前，中文文本自动分词技术有了很大的进步，以中文信息处理技术为基础的中文文本分类已趋于成熟，在短时间内经历了从可行性探索到应用性实践再到自动分类系统的发展历程。

1.2.2 特征选择的现状研究

上世纪 70 年代以来，众多学者已开始对特征选择的方法产生研究，W.Siedlecki 和 J.Sklansky[23-25]于 1988 年发表了一篇文章，提出了通过研究历史的时间来进行特征选择算法性能的评价，并讨论了 Branch&Bound 算法及其改进；J.Doak[26]研究了特征选择中的评价方法，即评价准则的问题；A.K.Jain 和 R.P.W.Duin[27]等人研究了搜索的起点、方向、策略问题；M.Dash 和 H.Liu[28]对以上方法进行了总结并根据文本选择的策略和评价标准进行分类。

根据特征选择过程中特征集中有无包含分类类标，可以将其分成有监督的和无监督的特征选择。

根据特征选择过程中是否有参与了分类算法，可将特征选择分为 Filter 方法、Wrapper 方法和 Embedded 方法三类。

Filter 方法是一种基于统计学、快速收敛、独立于学习过程、计算效率高、

对象是单一特征、以分析特征子集内部特点的特征选择方法。它是根据某些合适的判别准则[29-31]对特征属性进行评分，分值越高，特征越重要，再根据分值大小对特征进行排序和选择。目前使用最多的判别准则有信噪比指标（Signal to Noise Ratio，SNR）、t-statistic 指标、概率距离和相关测量法[29]、距离测量法[31,32]、信息熵法[33]等。基于信噪比采用加权投票的特征选择方法是典型的 Filter 方法。2000 年由 Arfin[34]和 Tanaka[35]提出的 t-statistic 方法有坚定的理论基础，但存在误发现率高的问题。Yang[36]等人为了降低误发现率，采用了多重假设检验相结合的策略。由于 Filter 方法基于单特征，各个特征之间的相互关系不高，会造成候选特征子集存在冗余特征，而且它不保证可以选出一个规模较小的良好特征子集。但其可以迅速剔除不重要的噪声，用于特征的预处理。

Wrapper 算法与其采用的分类模型紧密相关，在进行特征选择中，用以训练分类器，并且按照其在测试集上的表现作为选择标准，例如分类准确率。Wrapper 方法一般由候选特征集的搜索算法和分类器对候选特征集的学习算法两部分组成。搜索方法就是用来不断的产生候选的特征集，分类器对其的学习就是对各个特征集进行评估。遗传算法、前项序列选择（Sequential Forward Selection，SFS）和后项序列选择（Sequential Backward Selection，SBS）是我们比较常用的搜索算法。W.H.Hsu[37]提出了一种 GA 和 DT 相结合的特征选算法，来找出决策树分类中错误率最小的一个特征集；L.H.Chiang 和 R.J.Pell[38]提出了一种基于 Fisher 判别分析的特征选择方法，文中用遗传算法在识别化工故障里的关键变量的过程中获得不错的效果；I.Tabus 和 J.Astola[39]利用正太极大似然模型对特征集开展选择和分类，获得了很好的效果。常用的分类器包括支持向量机（SVM）、决策树（Decision Tree）[40-43]、人工神经网络（Artificial Neural Network，ANN）[44-48]、K 近邻[49-53]、Fisher 线性判别和贝叶斯分类器等。由于 Wrapper 方法与分类器有关，不同的分类器基于的理论各有不同，因而采用的特征重点也各异，所以在使用当前分类器得到的特征子集性能优良，在其他分类器不一定能取得同样良好的结果。而且 Wrapper 方法计算复杂度高，易于产生过拟合现象。

Embedded 方法与分类器性能结合的最为紧密，它采用分类器特有的某种特性来选择特征，所以要选用同时具有分类和特征选择功能的分类器，如：支持向